Tri-modal speech: Audio-visual-tactile integration in speech perception

Donald Derrick,[1,a)] Doreen Hansmann,[2] and Catherine Theys[2]
[1]*New Zealand Institute of Language, Brain, and Behaviour, University of Canterbury, 20 Kirkwood Avenue, Upper Riccarton, Christchurch 8041, New Zealand*
[2]*School of Psychology, Speech and Hearing, University of Canterbury, 20 Kirkwood Avenue, Upper Riccarton, Christchurch 8041, New Zealand*

(Received 20 May 2019; revised 3 September 2019; accepted 25 October 2019; published online 25 November 2019)

Speech perception is a multi-sensory experience. Visual information enhances [Sumby and Pollack (1954). J. Acoust. Soc. Am. **25**, 212–215] and interferes [McGurk and MacDonald (1976). Nature **264**, 746–748] with speech perception. Similarly, tactile information, transmitted by puffs of air arriving at the skin and aligned with speech audio, alters [Gick and Derrick (2009). Nature **462**, 502–504] auditory speech perception in noise. It has also been shown that aero-tactile information influences visual speech perception when an auditory signal is absent [Derrick, Bicevskis, and Gick (2019a). Front. Commun. Lang. Sci. **3**(61), 1–11]. However, researchers have not yet identified the combined influence of aero-tactile, visual, and auditory information on speech perception. The effects of matching and mismatching visual and tactile speech on two-way forced-choice auditory syllable-in-noise classification tasks were tested. The results showed that both visual and tactile information altered the signal-to-noise threshold for accurate identification of auditory signals. Similar to previous studies, the visual component has a strong influence on auditory syllable-in-noise identification, as evidenced by a 28.04 dB improvement in SNR between matching and mismatching visual stimulus presentations. In comparison, the tactile component had a small influence resulting in a 1.58 dB SNR match-mismatch range. The effects of both the audio and tactile information were shown to be additive. https://doi.org/10.1121/1.5134064

[BVT] Pages: 3495–3504

I. INTRODUCTION

We use speech every day and for much of our interpersonal communication, making accurate speech perception important to human interactions and well-being. Those with normal hearing can understand clear speech through the auditory signal alone, but speech perception is not a unimodal auditory-only process. We hear speech and can also see and feel it. Visual information enhances auditory speech perception of syllables, words, and phrases (Sumby and Pollack, 1954). Similarly, air flow hitting the skin in time with relevant speech audio (aero-tactile information) can enhance perception of degraded auditory syllables (Gick and Derrick, 2009) and words (Derrick *et al.*, 2019b), even in untrained and unaware perceivers. In addition, touching a speaker's face (Bernstein *et al.*, 1991) and feeling speech air flow (Bicevskis *et al.*, 2016) influences visual speech perception, absent any auditory information. These bi-modal audio-visual, audio-tactile, and visual-tactile interactions will be discussed in more detail below.

However, we do not yet know whether all three sensory modalities stack, let alone interact, to influence speech perception. Nor do we know the relative importance of visual stimuli as compared to tactile stimuli when the auditory signal is degraded. Here, we present a study to answer these questions during two-way forced-choice classification of audio, audio-tactile, and audio-visual-tactile syllables in speech noise.

A. Audio-visual speech perception

The first and still most-cited systematic effort to study audio-visual speech perception was conducted by Sumby and Pollack (1954). The Sumby and Pollack study predates the existence of effective technology for millisecond synchronization of audio and video. As a result, they used a live speaker speaking into a microphone, and digitally mixed the audio, scaling its volume to mix with noise in real-time and at the desired signal-to-noise ratio. The perceivers listened to that combined audio and noise through tight-fitting headphones. Participants in the audio-visual condition watched the speaker talk, while in the audio only condition, they faced away from the speaker, presumably looking at no face at all. Of note, the researchers were concerned about the possible influence of speech air flow, as indicated by asking participants to hold a cushion in front of their faces and bodies during the experiment to minimize air flow effects.

Sumby and Pollack (1954) showed that the importance of the visual component of speech is significant, and that its influence is strongest in noisier conditions and with simpler word identification tasks. When there was a choice between only eight words, perceivers reached over 90% accuracy in word identification at −30 decibels signal-to-noise ratio (SNR) for audio-visual speech, but only about 20% accuracy in audio-only speech. When participants could choose between

a)Electronic mail: donald.derrick@canterbury.ac.nz

128 and 256 words, their audio-visual accuracy dropped to about 40%, and the audio-only accuracy dropped to about 5% at the same SNR.

For decades, there was little follow-up research. One replication study (Erber, 1969), used television (Broadbent, 1970) instead of live faces, but replicated the Sumby and Pollack (1954) use of complex multi-forced-choice paradigms. Another study replicated Sumby and Pollack's study across aging populations (Ewertsen and Nielsen, 1971). During this time, it was still common to have participants in audio-visual experiments simply look away from the visual stimuli during audio conditions. In contrast, more recent studies, including those focusing on brain processing, use a still face for audio-only conditions when comparing them to audio-visual ones in order to achieve a more comparable baseline (e.g., Huyse *et al.*, 2014; Sekiyama *et al.*, 2003).

After the studies with matching audio-visual stimuli, McGurk and MacDonald (1976) used the first mismatched audio and visual stimuli in research. While they used the same look-away method for their auditory-only condition, they simplified research to open choice paradigms with two source stimuli (i.e., auditory "ba" and visual "ga"). They showed that mismatched audio and visual stimuli generate very different perceptual results compared to matching stimuli. Perceivers who listened to /ba-ba/ and watched a face saying /ga-ga/ would report hearing /da-da/ (McGurk and MacDonald, 1976) — the "McGurk effect." This fusion, however, is not the norm. Instead, perceivers usually experience a confusion; for instance when they heard /ga-ga/ and saw /ba-ba/, they often reported hearing combinations like /gabga/, or other structures like /dabda/. In limited-choice experiments, mismatched stimuli increase the SNR required for correct auditory identification (Mallick *et al.*, 2015; Sekiyama *et al.*, 2014).

Mismatched stimuli also alter the temporal windows of integration—in this case how closely aligned in time the audio and visual signal need to be for perceivers to pair them together. People have an easier time integrating audio-visual speech with delayed audio compared to delayed video (Dixon and Spitz, 1980; Smeele *et al.*, 1992; Summerfield, 1992). Munhall *et al.* (1996) show that audio information can precede visual information by 60 ms but follow it by as much as 240 ms. However, this asynchrony holds true most effectively for matching speech information. Audio-visual mismatches have narrower and more symmetrical windows of temporal integration, with about a 67 ms audio lag preferred (van Wassenhove *et al.*, 2007).

In more recent times, the number of audio-visual speech research papers has increased exponentially. There are now hundreds of articles examining the influence of audio-visual speech on behaviour and brain processing. These modern studies inform on audio-visual responses to matching and mismatching stimuli (Kaganovich *et al.*, 2016; Sekiyama *et al.*, 2003), temporal alignment of audio-visual stimuli (van Wassenhove *et al.*, 2007), and influences of audio-visual stimuli on brain processing, regions of brain activity for auditory and visual speech (Peelle and Sommers, 2015; van Wassenhove, 2013; Venezia *et al.*, 2016), audio-visual processing across childhood development (Ross *et al.*, 2011), the influence of disorders on audio-visual speech perception (Desai and Zen, 2008; Feldman *et al.*, 2018), and the differences between the perception of speech and non-speech (Tremblay *et al.*, 2007). Overall, the results tend to show that audio-visual speech perception interaction is strongest with young, non-visually or auditorily impaired, neurotypical populations, and when visual information aligns with or slightly precedes auditory information — matching typical experiences in live close-up speech.

B. Audio-tactile speech perception

Even before there was audio-visual speech research, clinicians working with hearing-impaired perceivers knew that feeling the speaker's face deformation, vibration, and air flow could enhance speech perception (Alcorn, 1932). Efforts to use vibro-tactile systems to encode speech (Rothenberg *et al.*, 1977) and to enhance face-reading (Bernstein *et al.*, 1991) were successful, though they required extensive participant training to be effective.

Later, researchers exploited a well-known feature of the speech of many languages (see Draper *et al.*, 1960; Lisker, 1957; Stathopoulos and Weismer, 1985; Tathan and Morton, 1973) — a categorizable distinction between high and low stop-release air flow in speech. If a native English speaker puts their hand in front of their face while saying "ba," they will experience low enough air flow and pressure to be unlikely to feel it, but if they do the same with "pa," they will experience a very noticeable air pop. Anyone who has ever made an audio recording or worked in radio knows this microphone pop can ruin a broadcast — which is why they use "pop-screens" in front of microphones or place the mic to the side of the speaker. This difference is best captured not through measures of intraoral air pressure, which varies surprisingly little between "ba" and "pa" (Tathan and Morton, 1973). Instead, voice onset time, which is a measure of frication duration and much lower for "ba" (average 1 ms), than "pa" (average 58 ms), is more informative (Lisker and Abramson, 1966).

That distinction in air flow can be produced artificially, aligned with audio in noise, and used to alter speech perception. Such tactile information enhanced accuracy of two-alternative forced-choice (2AFC) identification of voiceless stop onset auditory speech-in-noise syllables at the same SNRs from about 62.6% to 74.6% (12% range) when applied to the hand, and from 68.6% to 76.9% (8.3% range) when applied to the suprasternal notch (neck) (Gick and Derrick, 2009), and from 64.6% to 66.0% (1.4% range) when applied to the ankle (Derrick and Gick, 2013).

This type of air flow has also been shown to have a stronger effect on 2AFC syllable and word identification when the underlying differences in air flow for the two choices are greater (Derrick *et al.*, 2014). Air flow will help distinguish voiceless stops and some fricatives from voiceless stops but confuses people when presented for two sounds with similar air flow patterns. The effect appears to interact with the signal-to-noise ratio of the auditory signal such that the clearer the audio, the greater the required difference in underlying air flow for the two choices to result in

enhanced perception (Derrick *et al.*, 2019b). For instance, with Mandarin, when the air flow differences are close to zero, air flow can interfere with accurate perception, reducing accuracy from 65% to 56% (i.e., zh 扎 vs. ch 差), but when the air flow differences between two words are near the maximum for the language (i.e., d 搭 vs t 他), air flow can enhance accuracy from 55% to 65% (Derrick *et al.*, 2019b) at the same audio-in-noise SNR.

Mismatched tactile stimuli generate simple interference, such that perceivers are more likely to perceive voiced stops with simultaneous presentation of air puffs as voiceless (Derrick and Gick, 2013; Gick and Derrick, 2009; Gick *et al.*, 2010), or even as fricatives (Derrick *et al.*, 2014). In addition, matching auditory and tactile signals have asymmetrical windows of integration, whereas mismatching signals do not. Gick *et al.* (2010) showed that with matching signals, delayed air flow may be integrated as much as 100–200 ms after the relevant auditory signal, but delayed audio will only be integrated up to 50 ms after the air flow. However, the air flow must be within 50 ms of the audio signal for mismatched air flow to interfere with accurate syllable identification.

While studies of audio-tactile speech perception have shown that aero-tactile speech signals help to distinguish one auditory speech-in-noise signal from its alternative (Derrick and Gick, 2013; Derrick *et al.*, 2019c; Derrick *et al.*, 2014; Gick and Derrick, 2009; Gick *et al.*, 2010; Goldenberg *et al.*, 2015), these results have not extended to more complex tasks beyond 2AFC experiments (Derrick *et al.*, 2019b; Derrick *et al.*, 2014). Whether this represents a limited influence of aero-tactile information during complex speech perception, or instead is a reflection of the strength limitations of the aero-tactile stimuli used in those experiments, remains a topic for further research.

C. Visual-tactile speech perception

In addition to audio-visual and audio-tactile experiments, other bi-modal speech perception experiments removed the audio component of speech and asked participants to identify ambiguous (/ba/ vs /pa/) visual-tactile syllables. Participants identified but 37% of the combined /ba/ and /pa/ tokens as /pa/ in no-air flow conditions, but 65% /pa/ when the air flow was aligned to 100–150 ms after the burst onset (Bicevskis *et al.*, 2016). The effects of air flow were at their strongest when the air flow was aligned to start about 100–150 ms after the beginning of the lip opening, regardless of whether the visual stimuli was a /ba/ or a /pa/. Of note: These participants had narrower windows of integration the more they described themselves as neurotypical (Derrick *et al.*, 2019a), as measured through the Autism Spectrum Quotient Test (Baron-Cohen *et al.*, 2001). As has been done in audio-visual experiments, future studies should explore the factors that influence audio-tactile and visual-tactile integration in different populations.

D. Tri-modal speech perception

Audio-visual speech has been extensively studied, with results showing that matching visual speech can enhance and mismatching visual speech interfere with accurate auditory speech identification — with greater effects during simpler tasks and with temporally aligned speech signals. Though there are far fewer audio-tactile studies, similar results have been found in audio-tactile speech perception studies, and they have been replicated many times. Nevertheless, no research has been done that directly examines the relative importance of visual and tactile speech on auditory speech perception, let alone whether all three sensory modalities interact to influence speech perception.

While comparison of the relative improvements in SNR or accuracy across separate bi-modal studies suggest that tactile enhancement in auditory speech perception is likely less strong than visual enhancement of auditory speech perception, tactile stimuli have had considerable influence on auditory speech perception in two-way forced choice experiments. Previous bi-modal speech perception research had highly disparate complexities, ranging from 2-way forced-choice (Gick and Derrick, 2009) through to open-choice with two stimuli (McGurk and MacDonald, 1976), to 256-choice tasks (Sumby and Pollack, 1954), to the extremely complex tasks of Bernstein *et al.* (1991), where participants were asked to identify whole sentences in speech. This diversity, combined with lack of tri-modal testing, caused us to recognize a need to unify tri-modal speech in a simple, shared paradigm. Such a study should include mismatched as well as matched stimuli, with well aligned audio, visual, and tactile stimuli — within about 50 ms to fit the requirements for audio-tactile speech (Gick *et al.*, 2010), and within a minimally complex paradigm to maximize the potential influence of the tactile and visual signals.

Here, we present a two-way forced-choice experiment with 12 conditions representing two auditorily presented speech-in-noise syllables ([p^{h}a] and [ga]) along with matched and mismatched tactile (air flow vs no air flow) stimuli, along with matched, mismatched, and masked video of [p^{h}a] and [ga]. For each condition, the signal-to-noise ratio at 82% correct auditory signal identification will be identified using a fixed-response number QUEST staircase method (Watson, 1983). The accuracy threshold follows the QUEST staircase accuracy recommendations as documented in PsychToolBox3 (Kleiner *et al.*, 2007). The following hypotheses will be tested.

Hypothesis 1: Replicating previous experiments, visual stimuli will interact with audio stimuli to influence speech perception. The prediction is that the required audio SNRs for 82% syllable classification accuracy (based on the auditory signal) will be lower for matching video and audio stimuli (both [p^{h}a] or [ga]), and higher when the video and audio stimuli do not match (one [p^{h}a] and the other [ga]), with SNRs in between for cases where the video is masked.

Hypothesis 2: Replicating previous experiments, tactile stimuli will interact with audio stimuli to influence speech perception. The prediction is that required audio SNRs for 82% syllable classification accuracy (based on the auditory signal) will be lower for matching tactile and audio stimuli ([p^{h}a] with air flow, and [ga] without air flow), and higher when the tactile and audio stimuli do not match ([p^{h}a] without air flow, and [ga] with air flow).

Hypothesis 3: The effects of video and tactile stimuli will stack, but visual stimuli will have a larger influence. The prediction is that the required audio SNRs for 82% syllable classification accuracy will differ the most under the influence of visual and tactile stimuli (lowest with both matching and highest with both mismatching), followed by a weaker effect of visual stimuli, followed by the weakest, but still significant effect of tactile stimuli.

II. METHODS

The following methods describe the entire audio-visual-tactile closed-choice experiment. The still-video portion of these methods have been previously described in Derrick *et al.* (2019c), where those results are compared and contrasted to the results of a similarly-designed open-choice experiment. (In that open-choice experiment, air-flow did not significantly alter speech perception.)

A. Participants

The University of Canterbury Human Ethics Committee reviewed and approved this study, and participants provided informed consent. Participants then completed a demographic information sheet, reporting age, native language and history of speech, language and hearing difficulties. As part of the protocol, participants underwent an audiological screening. Pure tone audiometry testing was carried out for frequencies of 500 Hz, 1 kHz, 2 kHz, and 4 kHz using an Interacoustics AS608 screening audiometer. Average pure tone thresholds were calculated and if the threshold was less than or equal to 25 dB hearing level (HL), hearing sensitivity was considered to be within normal range. Forty (40) New Zealand English perceivers, 18–46 years old [$\mu = 24.6$, standard deviation (sd) = 8.0], 7 males, 33 female, then participated in this study.

B. Stimuli

1. Recording of Stimuli

One female speaker, producing forty tokens of "pa" [p^{h}a] and "ga" [ka] each, was recorded in a sound-attenuated room with a professional lighting setup. The video was recorded on a Sony MediaPro PMW-EX3 video camera set to record with the MPEG2 HD35 HL codec, with a resolution of 1920 by 1080 pixels (16:9 aspect ratio), a frame rate of 25 frames per second (fps), and a hardware-synched linear pulse-code-modulation (LPCM) 16-bit stereo audio recording at 48 000 Hz. The video was then converted to a time-preserving H.264 codec in yuv420p format encapsulated in an MP4 package, with audio extracted using FFMPEG (FFmpeg Developers, 2016). The audio was segmented in Praat (Boersma and Weenink, 2019), and the authors jointly selected ten recordings of each syllable that matched in duration, intensity, fundamental frequency, and phonation. In addition, the facial motion of each token was inspected to eliminate any case of eye-blink or noticeably distinguishable head motion.

2. Creation of A, AV, AT, and AVT stimuli

The ten "pa" and ten "ga" tokens were sorted by length to form the closest duration-matched pairs. Software was written in R (R Development Core Team, 2018), WarbleR (Araya-Salas and Smith-Vidaurre, 2017), FFmpeg (FFmpeg Developers, 2016), and the Macintosh Again Shell (BASH). The software took the timing of each video file and extracted the video with 750 ms lead time, and 500 ms follow time. For each video stimuli, it produced a version with right-channel audio from the original and left-channel audio that was either empty (for no air flow stimuli), or contained an 80 ms 12 kHz maximum intensity sine-wave used to operate our custom air flow system. In addition to the audio-visual (AV) condition, for each video, a version was produced with a blurred and still lower face for the audio only (A) condition, and a version with the audio from the paired alternative file (i.e., video "pa" with audio "ga," and video "ga" with audio "pa"), where the audio was aligned to the burst onset from the audio of the originating video for the audio-visual mismatch (AX) condition. This produced 12 types of stimuli, as seen in Table I, with a video appearance as seen in Fig. 1. Initial SNRs for each staircase were tuned from a pilot experiment of ten participants set up similarly to the one described here, but with poorer quality video (see Derrick *et al.*, 2016). These SNRs are also listed in Table I.

To generate speech noise, the recordings of the speech tokens were randomly superimposed 10 000 times within a 10 s looped sound file using an automated process written in R (R Development Core Team, 2018), WarbleR (Araya-Salas and Smith-Vidaurre, 2017), and FFMPEG (FFmpeg Developers, 2016). Noise created using this method results in a noise spectrum that is nearly identical to the long-term spectrum of the speech tokens from that speaker (Jansen *et al.*, 2010; Smits *et al.*, 2004). This type of noise has similar efficacy regardless of the volume at which it is presented, allowing for the useful application of signal-to-noise ratios used in this experiment.

The software then overlayed the right channel audio with speech-noise, making a video file for each token with signal-

TABLE I. Stimuli types.

Number	Audio	Video	Air Flow	Initial SNR	Number	Audio	Video	Air Flow	Initial SNR
1	"pa"	"pa"	yes	−15	7	"ga"	"ga"	yes	−12
2	"pa"	"pa"	no	−15	8	"ga"	"ga"	no	−10
3	"pa"	masked	yes	−8	9	"ga"	masked	yes	−10
4	"pa"	masked	no	0	10	"ga"	masked	no	−8
5	"pa"	"ga"	yes	1	11	"ga"	"pa"	yes	−5
6	"pa"	"ga"	no	3	12	"ga"	"pa"	no	−4

FIG. 1. (Color online) Video just prior to release burst.

to-noise ratios from −30 dB to +15 dB, at 0.1 dB increments. The noise overlay was attenuated for all tokens above 0 dB, and the underlying audio was attenuated for tokens below 0 dB, ensuring that each token was of similar maximum amplitude for maximum comfort during the experiments.

3. Stimulus presentation

The 80 ms 12 kHz sine-wave was used to operate our air flow production system (Derrick and De Rybel, 2015). The air flow system uses a Murata's microblower, a 20 × 20 × 1.85 mm piezoelectric air pump with up to 0.8 l/m flow, max 19.38 cm/H2O pressure, and approximately 30 ms 5%–95% intensity rise time, allowing artificial approximation of continuously-varying air flow in speech (Derrick *et al.*, 2015). The sine-wave in the left channel turns on the air flow system at full capacity, generating its highest air flow with a duration within the range of the voice onset time of a word-onset velar voiceless stop ("ka"), and at the long end of length for that of a labial voiceless stop ("pa") (Lisker and Abramson, 1966). The audio in the right channel was presented in both ears via Panasonic RP-HT265 closed stereo headphones at a comfortable loudness level, simultaneous with the relevant video.

C. Experimental procedure

Once the initial screening protocol was completed, participants were seated in a sound-attenuated booth, with a screen behind glass positioned 1 meter from participants. The air puff system (Derrick and De Rybel, 2015) was positioned on a microphone boom arm 3 cm from the participants' suprasternal notch (at the base of the neck and above the sternum). The entire control system is schematized in Fig. 2.

Participants were told that they may experience some noise and unexpected puffs of air along with speech syllables. During the experiment, participants were presented with the auditory, visual, and tactile stimuli and asked to press a key to choose between one of two possible syllables, "pa" or "ga." The experiment presented 12 conditions interleaved into QUEST staircases with 40 tokens each, or 480 tokens total, taking about 20 min.

Two-alternative forced-choice (2AFC) QUEST adaptive staircases (Watson, 1983) were written in MATLAB (The MathWorks, Inc., 2014) using the Psych Toolbox 3 software tools (Brainard, 1997; Kleiner *et al.*, 2007; Pelli, 1997). The QUEST staircases were tuned to identify the 82% accuracy threshold, with 40 trials for each of the 12 randomly interleaved blocks listed in Table I. The QUEST staircases used the standard Weibull function steepness (3.5), standard granularity of 0.01 dB SNR, and a wide latitude for allowable standard deviation (20 dB SNR) as per the protocol recommended in the Psych Toolbox manual. After each run, the QUEST quantile results, rounded to the nearest 0.1 dB, were used for the selection of stimuli, with the QUEST mean result used for final analysis, as per the recommendation in Pelli (1987).

D. Analysis

SNRs at 82% accuracy were extracted for each of the conditions. Linear mixed-effects models (LMM) were run on the interaction between tactile and visual match vs mismatch. The *p*-values were generated using the lmerTest package (Kuznetsova *et al.*, 2017), which itself uses the lmer function for generating LMMs as part of the lme4 package (Bates *et al.*, 2015). *Tactile match* included the six staircases where "pa" tokens were paired with air flow, and "ga" tokens were not. The remaining six staircases were mismatches. *Video match* was an indication of whether the audio and video match lined up, with three possibilities: (1) matches (AV) were audio and video "pa" and audio and video "ga," (2) mismatches (AX) were audio "pa" and video "ga," as well as audio "ga" and video "pa," and (3) none (A), where the video has the lower half of the face blurred, as seen in Fig. 1. Model fitting was then performed in a stepwise backwards iterative fashion, beginning with the model shown in Eq. (1)

$$SNR \sim VideoMatch * TactileMatch + (1 + (VideoMatch * TactileMatch)|Participant). \quad (1)$$

Models were back-fit along the Akaike information criterion (AIC), to measure quality of fit. This technique allows for the isolation of a statistical model that provides the best fit for the data and allowed elimination of independent variables from models that did not contribute to the explanatory power of the overall model. The final model can be seen in Eq. (2). Note that in this final model, there were no

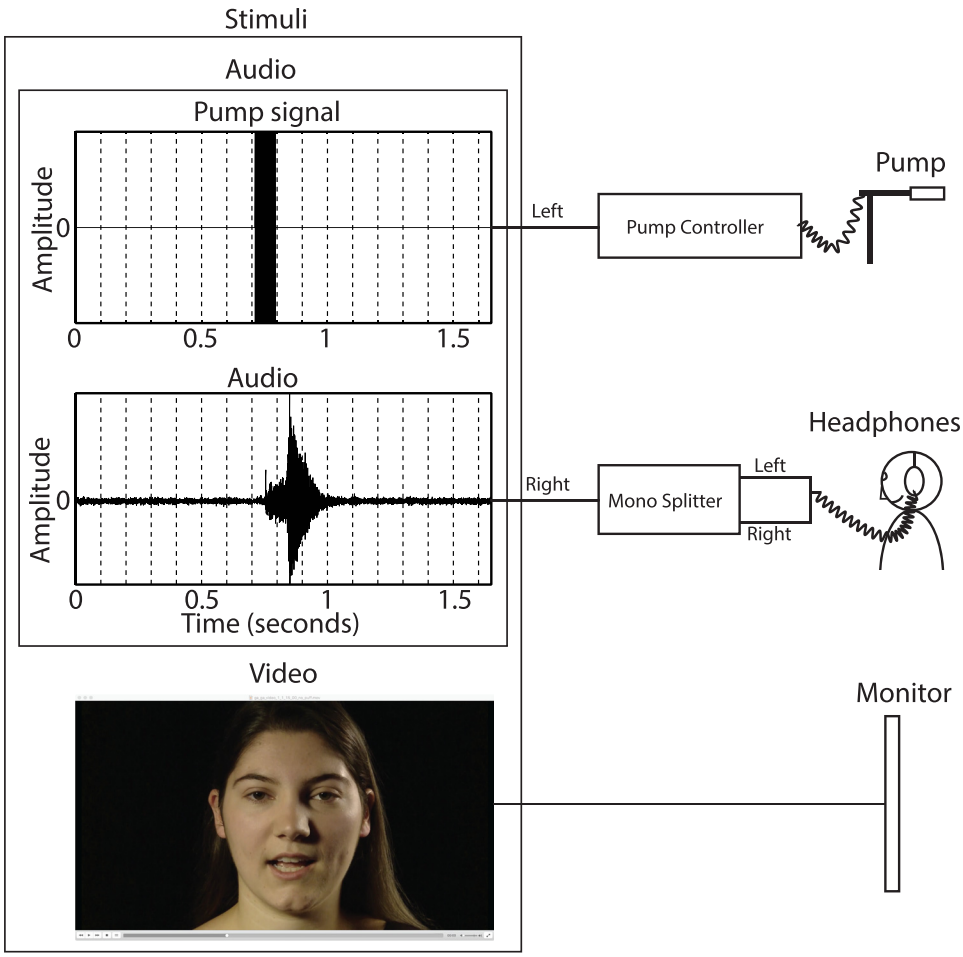

FIG. 2. (Color online) Experiment setup schematic. R = right channel audio, L = left channel audio, M = mono (right and left channel audio, V = video).

significant interactions between video and tactile matching parameters.

$$SNR \sim VideoMatch + TactileMatch + (1 + (VideoMatch + TactileMatch)|Participant). \quad (2)$$

III. RESULTS

Speech in noise levels at the 82% accuracy identification level of the auditory signal were identified for the 12 conditions. The results are first summarized in the descriptive results section, with sub-sections for the audio-visual and audio-tactile results, allowing comparison of the effect-sizes between the two. This analysis is followed by the overall statistical results, which highlights the best-fit analysis. The R-code used to obtain descriptive statistics, as well as the entire back-fitting process, is documented in the supplementary materials.[1]

A. Descriptive results

Figure 3 shows the notched boxplots of signal-to-noise ratio by audio-visual-tactile condition. It clearly shows the importance of the visual component in multi-modal speech, with lower SNRs for the AV compared to the A and AX

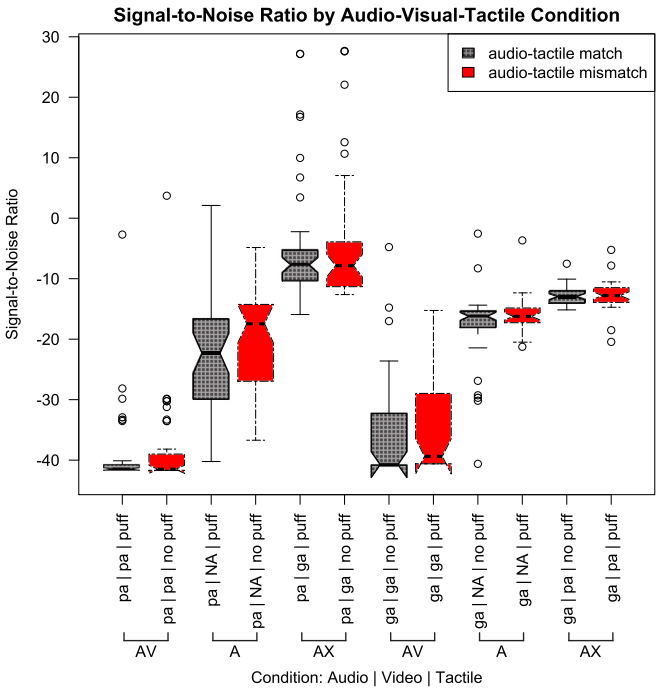

FIG. 3. (Color online) Notched boxplots of SNR by Audio-Visual-Tactile Condition. Each condition is listed by content as "audio | video | tactile" (e.g., pa | ga | puff = auditory "pa," visual "ga," and air flow). NA = no video, AV = audio-visual match, A = audio-only (still face), AX = audio-visual mismatch.

conditions. The effects of air flow on the 82% SNRs are less obvious compared to the visual effects but are still visible in all the staircase comparisons. The audio-tactile matched conditions, shown in hatched-grey, have lower SNRs than the mismatched conditions shown in red. Note that these boxplots are default R boxplots. The bars in the center of the rectangles are the median value, the notches represent the 95% confidence intervals of the median, the range of the rectangles represents the inter-quartile range (IQR), the whiskers represent 1.5 multipliers of the IQR, and circles represent values outside these ranges, sometimes called outliers.

When auditory and visual speech matched (AV), the auditory speech was accurately classified at an 82% threshold at −36.67 dB SNR [standard error (SE) 0.46 dB SNR]. In contrast, when the visual information was blurred (A), perceivers reached 82% accuracy at −19.44 dB SNR (SE 0.42 dB SNR). When the visual information was mismatched with the auditory information (AX), the 82% accuracy was lower at −8.63 dB SNR (SE 0.48 dB SNR). It is possible that the real SNR for AV could be even lower since the QUEST staircases hit floor effects for both matched and mismatched tactile stimuli for 13 of the 40 perceivers for AV "pa," and 4 of the 40 perceivers for AV "ga."

When air flow matched with the auditory stimuli, perceivers reached 82% accuracy at −22.37 dB SNR (SE 0.66 dB SNR), and when it mismatched, perceivers reached accuracy at −20.79 dB SNR (SE 0.63 dB SNR). These results are summarized in Table II.

B. Statistical results

The differences in SNRs between the conditions is statistically significant, as seen in the results of the linear mixed-effects model [Eq. (2)] shown in Table III. The estimates in the table present the ranges, all of which are obtainable from the descriptive statistics in Table II. Video match (AV) serves as the baseline, and the estimates show the difference in SNRs required for video mismatch (AX), which is 28.04 dB clearer audio, and for still video (A), which is 17.23 dB SNR clearer audio. Puff match also serves as a baseline, with puff mismatch requiring 1.58 dB SNR clear audio. As noted above, the best-fit model shows no interaction between visual and aero-tactile stimuli.

If we rerun the models with tokens showing floor effects across tactile match and mismatches for the AV condition, the results are very similar to those shown in Table III, but with a smaller AX vs AV range of 37.37 dB SNR, and A vs AV range of 16.44 dB SNR, and a slightly larger tactile mismatch vs match range of 1.67 dB SNR. Full details can be found in the supplementary materials.[1]

TABLE II. Mean, SD, and SE for response SNRs by tactile and video match.

Match Type	Mean	SD	SE
Video match	−36.67	8.27	0.65
Video mismatch	−8.62	8.56	0.68
Still video	−19.44	7.56	0.60
Tactile match	−22.37	14.41	0.93
Tactile mismatch	−20.79	13.83	0.89

TABLE III. Fixed-effects results of SNR differences (Observations: 480, Subjects: 40).

Fixed Effect	Estimate	Standard Error	df	*T*-value	*p*-value
(Intercept)	−37.46	1.15	40.37	−32.57	<0.001
Video mismatch (AX) vs match (AV)	28.04	1.32	39.08	21.25	<0.001
Still video (A) vs match (AV)	17.23	0.91	58.24	18.95	<0.001
Tactile mismatch vs match	1.58	0.62	287.68	2.55	0.011

IV. DISCUSSION

A large body of research has shown that speech perception is a multimodal process, taking not only auditory, but also visual and tactile information into account, especially when the auditory signal is degraded. Despite considerable evidence that speech perception can be a bi-modal audio-visual and audio-tactile process, with some support for visual-tactile speech perception, no study had shown the combined effect of perceiving all three modalities together.

This study combined all three modalities. The influence of the visual stimuli on auditory SNRs was very strong. During the auditory-only condition, the SNR at the 82% correct identification level was −19.44 dB SNR. In comparison, the average 82% threshold was −36.67 dB SNR when the visual information matched the auditory signal. This figure was obtained through the least-squares analysis from the QUEST algorithm (Watson, 1983), obtainable even though the lowest SNR presented to any participant was −30 dB. At the same time, when the visual information was mismatched, it interfered with auditory perception such that the average 82% accuracy threshold during audio-visual mismatch conditions was −8.63 dB SNR. Therefore, the range between the mismatching and matching audio-visual conditions was very large at 28.04 dB. In the auditory-only and the mismatching audio-visual conditions none of the participants demonstrated floor effects. In the matching audio-visual condition, floor effects for both tactile match and mismatch were present for 13 and 4 participants for "pa" and "ga," respectively. These participants did not answer incorrectly no matter how much noise obscured the auditory signal. This suggests that many of the perceivers relied on visual stimuli alone to identify strongly degraded auditory signals. Similar plateau effects due to reliance on visual information alone have been shown in previous audio-visual studies using matching stimuli (Sumby and Pollack, 1954).

In contrast, the influence of tactile stimuli altered the 82% accuracy signal-to-noise ratio by only 1.58 dB SNR. This is weaker, but still statistically significant with a *t*-value of 2.55. These results also show that visual and tactile information stack and so both influence speech perception at the same time, showing that speech perception is truly multimodal. When the models were reran with the floor effected responses removed, the 82% accuracy signal-to-noise range

for tactile match vs mismatch was only slightly higher at 1.67 dB SNR. These results suggest that floor effects may have obscured some of the statistical power of the tactile stimuli.

Our results pattern well with those from previous bimodal research. As noted, in audio-tactile research, tactile information enhanced accuracy of two-way forced-choice (2AFC) identification of voiceless stop onset syllables from about 68.6% to 76.9% (8.3% range) when applied to the suprasternal notch (neck) (Gick and Derrick, 2009). In our study, presenting matching aero-tactile stimuli at the suprasternal notch lead to a decrease in required auditory clarity by 1.58 dB SNR. Similarly, the larger 28.04 dB SNR effect seen in the video match-mismatch comparison matches the results from Sumby and Pollack's work. Sumby and Pollack (1954) show that adding visual information enhanced accuracy from about 65% to about 98% (33% range) at the same signal-to-noise ratio (-15 ± 2 dB SNR) with the least complex task they ran (8-way forced-choice). Sumby's visual results were less powerful in more complex word sets (up to 256-way forced-choice), so we expected greater power again in our two-way forced-choice experiments. This is further supported by Massaro's findings that speech stimuli are more often correctly reported in forced-choice than open-choice experiments (Massaro, 1998).

The relative difference between the influence of visual and tactile stimuli helps explain why we have observed different results for aero-tactile integration in speech perception based on task complexity. We observe strong aero-tactile integration in speech categorization during two alternative forced-choice (2AFC) tasks (Derrick and Gick, 2013; Gick and Derrick, 2009; Gick *et al.*, 2010). However, recent evidence goes against aero-tactile integration during more complex speech perception tasks (Derrick *et al.*, 2019c; Goldenberg *et al.*, 2018). It is possible that aero-tactile integration in speech perception can easily be overwhelmed by more influential auditory and visual information during complex speech.

In addition to the relative comparison between the influence of audio-visual and audio-tactile conditions, our study focused on the effect of combining all three modalities at once. The results of the matching and mismatching audio-visual-tactile conditions showed that both video and tactile stimuli can significantly enhance or interfere with auditory speech perception, and that they stack together into tri-modal effects on behavioural responses. Yet the three modalities did not interact. The lack of interaction between visual and tactile effects on speech perception may be a counter to earlier evidence in support of speaker preference for either visual or tactile integration (Gick *et al.*, 2008). Instead, the results show that visual and tactile stimuli additively stack in their influence on speech perception. In this study, visual plus tactile information combined to have a stronger influence on auditory speech perception than visual information alone, which itself had a stronger influence than tactile stimuli alone. To further test whether tri-modal effects occur in a simple additive manner rather than resulting from an interaction between the different types of sensory processing, functional brain imaging or electrophysiological methods could be used. Findings of such studies could help clarify whether parallel or integrative neural pathways are used in tri-modal speech perception. In addition, a full comparison of the effect of tri-modal speech on perception would require the ability to fully compare all matching and mismatching combinations. However, this comparison is not truly complete due to the floor effects reached when congruent audio and visual information was presented. Future studies using different levels of auditory, visual, and tactile degradation would allow identification of a true three-dimensional map of the strengths of auditory, visual, and tactile signal effects.

We kept the design as simple as possible so that we could achieve our desired results without putting undue burdens on our participants. Nevertheless, the experiment setup was still highly complex, pushing the boundaries of current technology. Those seeking to replicate or extend this study should go beyond the standard testing and piloting of their experiments. For example, simulated runs stress-testing the equipment could be beneficial to avoid intermittent failures that may result from the interaction of their computers, operating systems, audio and video codecs, and external components.

The simplicity of the design, while methodologically and technologically necessary, also resulted in limitations beyond the already mentioned floor effects. The results show high variability as indicated by the number of visible outliers, particularly for the A conditions for "pa," and the AV conditions for "ga," and the AX conditions for "pa." We do not know why this result occurred except that different participants had very different degrees of difficulty with the task. Recent results on visual-tactile integration in speech perception suggest that people have very different windows of integration during multisensory speech perception related to their self-reported autism spectrum quotients (Derrick *et al.*, 2019a).

Future studies would benefit from collection of data on neurotypicality, and on having more than two stimuli as each stimulus contains differing auditory, visual, and tactile information. The results here showed visual "ga" had more influence on perception of auditory "pa" than vice versa. We can therefore expect similar patterns of visual influence based on the amount of relevant information in the visual as compared to the auditory signal. The relationship between the amount of relevant information in each modality's signal is therefore highly worthy of future study.

To conclude, we presented the first systematic tri-modal audio-visual-aero-tactile speech perception experiment. It clearly demonstrates the simultaneous influence of both visual and tactile signals on auditory speech perception. It also shows that the relative importance of visual speech outweighs that of tactile speech, but that this weighting is modulated by the amount of information in the visual signal. Future studies using different levels of degradation of each of the sensory stimuli, as well as future studies of brain processing are necessary to create a full picture of tri-modal speech perception.

ACKNOWLEDGMENTS

This research was supported by New Zealand Ministry of Business, Innovation, and Employment (MBIE) Grants

"Aerotactile Enhancement in Speech Perception" and "Aerotactile Enhancement in Speech Perception — Phase II," and a University of Canterbury Marsden Support Fund grant to D.D. and C.T. (Co-PI's). Special thanks to John Chrisstoffels (University of Canterbury School of Fine Arts) for his cinematography, and to Claire Elliott for providing our stimuli. Special thanks to Jonathan Wiltshire for PsyToolBox support on Windows machines, and Scott Lloyd for his technical support of the air flow system.

[1]See supplementary material at https://doi.org/10.1121/1.5134064 for information on the R-code used to obtain descriptive statistics as well as the entire back-fitting process.

Alcorn, S. (**1932**). "The Tadoma method," Volta Rev. **34**, 195–198.
Araya-Salas, M., and Smith-Vidaurre, G. (**2017**). "warbleR: An R package to streamline analysis of animal acoustic signals," Methods Ecol. Evol. **8** (2), 184.
Baron-Cohen, S., Wheelwright, S., Skinner, R., Martin, J., and Clubley, E. (**2001**). "The autism-spectrum quotient (AQ): Evidence from Asperger syndrome/high-functioning autism, males and females, scientists and mathematicians," J. Autism Dev. Disorders **31**(1), 5–17.
Bates, D., Mächler, M., Bolker, B., and Walker, S. (**2015**). "Fitting linear mixed-effects models using lme4," J. Stat. Softw. **67**(1), 1–48.
Bernstein, L. E., Demorest, M. E., Coulter, D. C., and O'Connell, M. P. (**1991**). "Lipreading sentences with vibrotactile vocoders: Performance of normal-hearing and hearing-impaired subjects," J. Acoust. Soc. Am. **90**(6), 2971–2984.
Bicevskis, K., Derrick, D., and Gick, B. (**2016**). "Visual-tactile integration in speech perception: Evidence for modality neutral speech primitives," J. Acoust. Soc. Am. **140**(5), 3531–3539.
Boersma, P., and Weenink, D. (**2019**). "Praat: Doing phonetics by computer (version 6.0.52) [computer program]," http://www.praat.org (Last viewed May 2, 2019).
Brainard, D. H. (**1997**). "The psychophysics toolbox," Spatial Vis. **10**, 433–436.
Broadbent, D. (**1970**). "Televised visual contribution to speech recognition," IEEE Trans. Educ. **13**(2), 79–82.
Derrick, D., Bicevskis, K., and Gick, B. (**2019a**). "Visual-tactile speech perception and the autism quotient," Front. Commun. Lang. Sci. **3**(61), 1–11.
Derrick, D., and De Rybel, T. (**2015**). "System for audio analysis and perception enhancement," PCT patent no. WO 2015/122785 A1.
Derrick, D., De Rybel, T., and Fiasson, R. (**2015**). "Recording and reproducing speech airflow outside the mouth," Can. Acoust. **43**(3), 108–109, availabe at https://jcaa.caa-aca.ca/index.php/jcaa/issue/view/268/pdf_198.
Derrick, D., and Gick, B. (**2013**). "Aerotactile integration from distal skin stimuli," Multisens. Res. **26**, 405–416.
Derrick, D., Heyne, M., O'Beirne, G., and Hay, J. (**2019b**). "Aero-tactile integration in Mandarin," in *Proceedings of the 19th International Congress of Phonetic Sciences*, August 5–9, Melbourne, Australia, pp. 3508–3512.
Derrick, D., Madappallimattam, J., and Theys, C. (**2019c**). "Aero-tactile integration during speech perception: Effects of an open-choice task," J. Acoust. Soc. Am. **146**(3), 1605.
Derrick, D., O'Beirne, G. A., De Rybel, T., and Hay, J. (**2014**). "Aero-tactile integration in fricatives: Converting audio to air flow information for speech perception enhancement," in *Proceedings of the 15th Annual Conference of the International Speech Communication Association (INTERSPEECH 2014)*, September 14–18, Singapore, pp. 2580–2584.
Derrick, D., O'Beirne, G. A., Gordon, M., De Rybel, T., Fiasson, R., and Hay, J. (**2016**). "Effects of aero-tactile stimuli on continuous speech perception," in *5th Joint Meeting, Acoustical Society of America and Acoustical Society of Japan*, November 28–December 2, Honolulu, HI.
Desai, S., and Zen, G. S. F.-G. (**2008**). "Auditory-visual speech perception in normal-hearing and cochlear-implant listeners," J. Acoust. Soc. Am. **123**(1), 428–440.
Dixon, N., and Spitz, L. (**1980**). "The detection of audiovisual desynchrony," Perception **9**, 719–721.
Draper, M. H., Ladefoged, P., and Whitteridge, D. (**1960**). "Expiratory pressures and air flow during speech," Br. Med. J. **18**(5189), 1837–1843.
Erber, N. (**1969**). "Interaction of audition and vision in the recognition of oral speech stimuli," J. Speech Hear. Res. **12**, 423–425.
Ewertsen, H., and Nielsen, H. B. (**1971**). "A comparative analysis of the audiovisual, auditive and visual perception of speech," Acta Otolaryngol. **72**, 201–205.
Feldman, J. I., Dunham, K., Cassidy, M., Wallace, M. T., Liu, Y., and Woynaroski, T. G. (**2018**). "Audiovisual multisensory integration in individuals with autism spectrum disorder: A systematic review and meta-analysis," Neurosci. Biobehav. Rev. **95**, 220–234.
FFmpeg Developers (**2016**). "FFmpeg tool [sofware]," http://ffmpeg.org/ (Last viewed May 2, 2019).
Gick, B., and Derrick, D. (**2009**). "Aero-tactile integration in speech perception," Nature **462**, 502–504.
Gick, B., Ikegami, Y., and Derrick, D. (**2010**). "The temporal window of audio-tactile integration in speech perception," J. Acoust. Soc. Am. **128** (5), EL342–EL346.
Gick, B., Jœhannsdœttir, K. M., Gibraiel, D., and Mhlbauer, J. (**2008**). "Tactile enhancement of auditory and visual speech perception in untrained perceivers," J. Acoust. Soc. Am. **123**(4), EL72–EL76.
Goldenberg, D., Tiede, M., and Whalen, D. H. (**2018**). "Concurrent aero-tactile stimulation does not bias perception of vot for non-initial stops," J. Acoust. Soc. Am. **144**(3), 1801–1801.
Goldenberg, D., Tiede, M. K., and Whalen, D. H. (**2015**). "Aero-tactile influence on speech perception of voicing continua," in *Proceedings of the 18th International Congress of the Phonetic Sciences (ICPhS2015)*, August 10–14, Glasgow, Scotland.
Huyse, A., Leybaert, J., and Berthommier, F. (**2014**). "Effects of aging on audio-visual speech integration," J. Acoust. Soc. Am. **136**(4), 1918–1931.
Jansen, S., Luts, H., Wagener, K. C., Frachet, B., and Wouters, J. (**2010**). "The French digit triplet test: A hearing screening tool for speech intelligibility in noise," Int. J. Audiol. **49**(5), 378–387.
Kaganovich, N., Schumaker, J., and Rowland, C. (**2016**). "Matching heard and seen speech: An ERP study of audiovisual word recognition," Brain Lang. **157–158**, 14–24.
Kleiner, M., Brainard, D., and Pelli, D. (**2007**). "What's new in psychtoolbox-3?," in *Perception Thirtieth European Conference on Visual Perception Abstract Supplement*, August 27–31, Arezzo, Italy.
Kuznetsova, A., Brockhoff, P. B., and Christensen, R. H. B. (**2017**). "lmerTest package: Tests in linear mixed effects models," J. Stat. Softw. **82**(13), 1–26.
Lisker, L. (**1957**). "Closure duration and the intervocalic voiced-voiceless distinction in English," Ling. Soc. Am. **33**(1), 42–49.
Lisker, L., and Abramson, A. S. (**1966**). "Some effects of context on voice onset time in English stops," Lang. Speech **10**, 1–28.
Mallick, D. B., Magnotti, J. F., and Beauchamp, M. S. (**2015**). "Variability and stability in the McGurk effect: Contributions of participants, stimuli, time, and response type," Psychonom. Bull. Rev. **22**(5), 1299–1300.
Massaro, D. W. (**1998**). *Perceiving Talking Faces: From Speech Perception to a Behavioural Principle* (MIT Press, Cambridge, MA), Vol. 1.
McGurk, H., and MacDonald, J. (**1976**). "Hearing lips and seeing voices," Nature **264**, 746–748.
Munhall, K. G., Gribble, P., Sacco, L., and Ward, M. (**1996**). "Temporal constraints on the McGurk effect," Atten. Percept. Psychophys. **58**(3), 351–362.
Peelle, J. E., and Sommers, M. S. (**2015**). "Prediction and constraint in audiovisual speech perception," Cortex **68**, 169–181.
Pelli, D. G. (**1987**). "The ideal psychometric procedure," Investig. Opthalmol. Visual Sci. **20**, 366.
Pelli, D. G. (**1997**). "The videotoolbox software for visual psychophysics: Transforming numbers into movies," Spatial Vis. **10**, 437–442.
R Development Core Team (**2018**). *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna, Austria).
Ross, L. A., Molholm, S., Blanco, D., Gomez-Ramirez, M., Saint-Amour, D., and Foxe, J. J. (**2011**). "The development of multisensory speech perception continues into the late childhood years," Eur. J. Neurosci. **33**, 2329–2337.
Rothenberg, M., Verillo, R. T., Zahorian, S. A., Brachman, M. L., and Bolanowski, S. J., Jr. (**1977**). "Vibrotactile frequency for encoding a speech parameter," J. Acoust. Soc. Am. **62**(4), 1003–1012.
Sekiyama, K., Kanno, I., Miura, S., and Sugita, Y. (**2003**). "Auditory-visual speech perception examined by fMRI and PET," Neurosci. Res. **47**, 277–287.
Sekiyama, K., Soshi, T., and Sakamoto, S. (**2014**). "Enhanced audiovisual integration with aging in speech perception: A heightened mcgurk effect in older adults," Front. Psychol. **5**(323), 1–12.

Smeele, P. M. T., Sittig, A. C., and Heuven, V. J. V. (**1992**). "Intelligibility of audio-visually desynchronized speech: Asymmetrical effect of phoneme position," in *Proceedings of the International Conference on Spoken Language Processing*, October 13–16, Alberta, Canada, pp. 65–68.
Smits, C., Kapteyn, T. S., and Houtgast, T. (**2004**). "Development and validation of an automatic speech-in-noise screening test by telephone," Int. J. Audiol. **43**(1), 15–28.
Stathopoulos, E. T., and Weismer, G. (**1985**). "Oral airflow and air pressure during speech production: A comparative study of children, youths and adults," Folia Phon. Logopaed. **37**(3–4), 152–159.
Sumby, W. H., and Pollack, I. (**1954**). "Visual contribution to speech intelligibility in noise," J. Acoust. Soc. Am. **26**, 212–215.
Summerfield, Q. (**1992**). "Lipreading and audio-visual speech perception," Philos. Trans. R. Soc. Lond. Ser. B **335**, 71–78.
Tathan, M., and Morton, K. (**1973**). "Electromyographic and intraoral air pressure studies of bilabial stops," Lang. Speech **16**(4), 336–350.
The MathWorks, Inc. (**2014**). *MATLAB and Statistics Toolbox Release 2014b* (The MathWorks, Inc., Natick, MA).
Tremblay, C., Champoux, F., Voss, P., Bacon, B. A., Lepore, F., and Théoret, H. (**2007**). "Speech and non-speech audio-visual illusions: A developmental study," PLoS ONE **8**, e742.
van Wassenhove, V. (**2013**). "Speech through the ears and eyes: Interfacing the senses with the supramodal brain," Front. Psychol. **4**(368), 1–17.
van Wassenhove, V., Grant, K. W., and Poeppel, D. (**2007**). "Temporal window of integration in auditory-visual speech perception," Neuropsychologia **45**, 598–607.
Venezia, J. H., Thurman, S. M., Matchin, W., George, S. E., and Hickok, G. (**2016**). "Timing in audiovisual speech perception: A mini review and new psychophysical data," Atten. Percept. Psychophys. **78**(2), 583–601.
Watson, A. B. (**1983**). "QUEST: A Bayesian adaptive psychometric method," Percept. Psychophys. **33**(2), 113–120.